Extractive Summarization of Legal Decisions using Multi-task Learning and Maximal Marginal Relevance

Abhishek Agarwal and **Shanshan Xu** and **Matthias Grabmair**
Technical University of Munich, Germany
{abhishek.agarwal, shanshan.xu, matthias.grabmair}@tum.de

Abstract

Summarizing legal decisions requires the expertise of law practitioners, which is both time- and cost-intensive. This paper presents techniques for extractive summarization of legal decisions in a low-resource setting using limited expert annotated data. We test a set of models that locate relevant content using a sequential model and tackle redundancy by leveraging maximal marginal relevance to compose summaries. We also demonstrate an implicit approach to help train our proposed models generate more informative summaries. Our multi-task learning model variant leverages rhetorical role identification as an auxiliary task to further improve the summarizer. We perform extensive experiments on datasets containing legal decisions from the US Board of Veterans' Appeals and conduct quantitative and expert-ranked evaluations of our models. Our results show that the proposed approaches can achieve ROUGE scores vis-à-vis expert extracted summaries that match those achieved by inter-annotator comparison.

1 Introduction

In common-law systems, law practitioners research large numbers of legal decisions from past cases to find similar precedents that justify their arguments and lead to favorable outcomes. The analysis can be time-consuming and expensive as these documents are long and verbose, and understanding them requires legal expertise. Automatic summarization of legal documents can help expedite the process cost-effectively. However, the limited availability of expert-annotated summaries makes it challenging to design such automated systems to assist paralegals, lawyers, and other law practitioners.

Extractive summarization aims to identify and extract essential sentences from the source document to compose the corresponding summary. It is more common in the legal domain due to the complexity of the legal language and the scarcity of labeled data. By contrast, abstractive summarization generates an abstract representation that captures the salient ideas of the source text and might contain new words and phrases not present in the source document.

One of the main challenges of extractive summarization is the redundancy in legal documents, as legal decisions can often contain several semantically similar sentences. Our objective is to generate summaries that provide maximum information while minimizing redundancy. Maximal Marginal Relevance (*MMR*) (Carbonell and Goldstein, 1998) has proved to be an effective tool to tackle redundancy explicitly (Zhong et al., 2019) by balancing the importance of query relevance and diversity. However, more recent methods like *MMR-Select* can use neural models as a substitute for query relevance. Additionally, we can train the neural models to handle the redundancy implicitly by adding a redundancy loss term (Xiao and Carenini, 2020).

Another challenge is the low availability of expert annotated summarization datasets in the legal domain. In this work, we leverage large amounts of unlabeled data along with the small annotated datasets to gain maximum performance. Pre-trained transformers like BERT (Devlin et al., 2019) can improve the performance of downstream tasks, such as summarization, even with limited labeled data. However, such models trained on the general domain may fail to capture the intricacies of the domain-specific vocabulary used in legal decisions. The domain-specific variants of BERT (Chalkidis et al., 2020; Zheng et al., 2021) pre-trained on large corpora of legal texts can help better embed the legal terms and achieve robust performance in various legal-specific downstream tasks like argument mining (Xu et al., 2021), rhetorical role labeling (Bhattacharya et al., 2021a), and legal citation recommendation (Huang et al., 2021).

To maximize the summarization performance, we also leverage Multi-task Learning (MTL) by

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1857 - 1872

aggregating training samples from several smaller datasets of multiple related tasks. MTL helps the model learn shared representations between the primary task (summarization) and the auxiliary task (rhetorical role identification) to generalize better. The identification of rhetorical roles involves identifying the function of different sentences to understand underlying reasoning and argument patterns in legal decisions. Previous works have often used rhetorical role labeling as a precursor to extractive summarization to improve performance (Zhong et al., 2019; Bhattacharya et al., 2021b). In this paper, we explore the idea of using rhetorical role identification as an auxiliary task to augment our annotated dataset and help generate better summaries.

In brief, we consider our contributions to the extractive summarization of legal documents as follows:

- We generate informative summaries with maximum information and minimum redundancy in a low-resource setting. Our experiments demonstrate a general improvement in ROUGE scores for the proposed approaches.
- We further improve the summarizer using a multi-task setting by combining extractive summarization and rhetorical role labeling. The quantitative evaluation demonstrates that the multi-task models perform better than the single-task models.
- We evaluate the generated summaries qualitatively with the help of a legal expert. In contrast to the quantitative evaluation, the qualitative results show that our proposed approaches rank at least as good as human annotators.[1]

2 Related Work

2.1 Extractive Summarization

Galgani et al. (2012) developed a rule-based approach to summarization that uses a knowledge base, statistical information, and other handcrafted features like POS tags, specific legal terms, and citations. Kim et al. (2012) propose a graph-based summarization system that constructs a directed graph for each document where nodes are assigned weights based on how likely words in a given sentence appear in the conclusion of judgments. CaseSummarizer (Polsley et al., 2016), an automated text summarization tool, uses word frequency augmented with additional domain-specific knowledge to score the sentences in the case document. Liu and Chen (2019) propose a classification-based approach that uses several handcrafted features as input. However, such techniques require knowledge engineering of different features and do not tackle redundancy in legal decisions. Recently, various proposed approaches have tried to address redundancy in legal decisions for purposes of summarization. Zhong et al. (2019) hypothesize that the iterative selection of predictive sentences using a CNN-based train-attribute-mask pipeline followed by a Random Forest classifier to distinguish between sentences containing Reasoning/EvidentialSupport and other types. *MMR* then selects the final sentences for the summary. (Bhattacharya et al., 2021b) demonstrate an unsupervised approach named DELSumm that generates extractive summaries by incorporating guidelines from legal experts into an optimization problem that maximizes the informativeness and content words, as well as conciseness. In this work, we use an *MMR*-based variant which tackles redundancy explicitly and can be combined with a neural classifier to generate summaries. It alleviates the need to engineer handcrafted features or specific expert guidelines to prevent redundancy.

2.2 Rhetorical Role Labeling

Saravanan and Ravindran (2010) propose a rule-based system along with a Conditional Random Field (CRF) approach to identify the different segments. Nejadgholi et al. (2017) proposed a semi-supervised approach to searching legal facts in immigration-specific case documents by using an unsupervised word embedding model to aid the training of a supervised fact-detecting classifier using a small set of annotated sentences. The authors in (Walker et al., 2019) compare the performance between rule-based scripts and ML algorithms to classify sentences that state findings of fact. Bhattacharya et al. (2019) explore the use of hierarchical BiLSTM models by adding an attention layer and experiment with the pre-trained word and sentence embeddings (Bhattacharya et al., 2021a). (Savelka et al., 2021) annotated legal cases from seven countries in six languages using a structural type system and found that Bi-GRU models could be generalized for data across different jurisdictions to some degree. Despite copious work, there are very few annotated rhetorical role datasets in the legal domain. In this work, we use rhetorical role label-

[1] Our code is available here

ing as an auxiliary task to augment our annotated dataset and help generate better summaries.

3 Data

We use the dataset containing single-issue Post-Traumatic Stress Disorder decisions from the US Board of Veterans' Appeals[2] (BVA) by (Zhong et al., 2019). These cases focus on veterans' appeals for benefits for a PTSD disability connected to stressful experiences during military service. The dataset is a sample from the BVA database that has been constrained to single-issue cases focusing on PTSD. In the texts, the BVA reviews the available evidence and either makes a finding that it warrants an award for service-connected PTSD (granted) or not (denied), or refers the case back to a lower administrative division for further development (remand). The dataset consists of 112 decisions and the corresponding expert annotated gold-standard summaries. We have 92 cases (48 remanded, 28 denied, 16 granted) in the training set with one annotated summary each. Another 20 cases (10 remanded, 6 denied, 4 granted) constitute the test set, for which there are four extractive summaries by different annotators and two drafted abstractive summaries. Each annotator chose a 6-10 sentence long summary based on predefined guidelines, out of which they selected 1-3 sentences each from the *Reasoning* and *Evidence* annotation type. The *Reasoning* sentences connect the outcome to the facts, while *Evidence* sentences add more information to support the former.

For rhetorical role labeling, we use two different datasets containing 50 plus 25 annotated BVA decisions[3](Walker et al., 2019). The decisions in the larger dataset have partial annotations, so we keep only decisions that have annotations for at least 60%[4] of the sentences in each decision. It results in 28 decisions consisting of 17 denied and 11 granted outcomes, while the smaller dataset contains 10 denied and 15 granted decisions. We map the different annotation types of the two datasets to a uniform type system of six annotation types. We merged the different annotation types for our experiments, resulting in 1889 *Evidence/Reasoning* and 3728 *Others* sentences. Therefore, the final dataset has 53 decisions with 7473 binary sentence-level annotations.

The datasets are from different time periods; therefore, the decisions have a slightly different document structure. We remove the meta-information like the case number, dates, judge names, names of the witnesses, and other similar information to have a more uniform layout. We keep only the following sections (if present) from each decision in the dataset:

- *Order*
- *Finding of Fact*
- *Conclusion of Law*
- *Reasons and Bases for Finding and Conclusion*
- *Remand*
- *Reasons for Remand*

Additionally, we use the SpaCy[5] pipeline enhanced with additional handcrafted rules to segment the sentences in each document for the summarization dataset.[6] After pre-processing, the average number of sentences per decision in the summarization and rhetorical role labeling datasets is **77.29 (± 52.28)** and **118.37 (± 78.33)**, respectively.

4 Our Approach

4.1 Sentence Embeddings

Sentence embeddings map an input sentence to a fixed-size dense vector representation. Sentence-BERT (Reimers and Gurevych, 2019) has recently emerged as an effective tool to derive semantically meaningful sentence embeddings. However, the lack of a domain-specific labeled entailment dataset required to train it makes it inaccessible for us. Alternatively, we can use a BERT model to extract a sentence embedding by pooling the embeddings for each token in the sentence. The mean-pooling operation outputs a 768-dimensional fixed-sized representation for each sentence. Such embeddings generalize quite well and provide a good starting point for training our sequential models later. This work uses the legal-domain specific transformer Legal-BERT (Zheng et al., 2021) trained on the Harvard Law case corpus' 3,446,187 legal decisions to generate the sentence embeddings.

[2] https://www.bva.va.gov

[3] https://github.com/LLTLab/VetClaims-JSON

[4] The remaining sentences with missing annotations contain sentences from annotation types other than *Evidence* or *Reasoning*, so we automatically annotate these sentences with the *Others* type.

[5] https://spacy.io

[6] To validate the performance of the sentence segmentation, we manually segment 11 decisions (6 remanded, 3 denied, 2 granted) and compare the matches. In terms of Recall, our approach and the legal text sentence segmenter (Savelka et al., 2017) score **0.937** and **0.905**, respectively.

4.2 Weighted Loss Function

The conventional cross-entropy loss function for the extractive summarization results in poor classification performance due to the class imbalance. For each decision, on average, we have very few positive labels (5-6 sentences) for each summary, which results in a highly imbalanced dataset. To tackle this issue, we use the weighted cross-entropy loss function that puts more emphasis on positive labels by manually rescaling the weights for each class.

$$w_c = \frac{\#samples}{\#classes \times \#samples_c}$$

$$L_{CE} = -\sum_{c=1}^{M} w_c \, (y_{o,c} \log(p_{o,c}))$$

4.3 Maximal Marginal Relevance

Maximal Marginal Relevance (*MMR*) (Carbonell and Goldstein, 1998) iteratively (greedily) selects sentences for the summary while balancing the query relevance and diversity:

$$\text{MMR} = arg \max_{s_i \epsilon D \setminus \hat{S}} [\, \lambda \, Sim(s_i, Q) - (1-\lambda) \max_{s_j \epsilon \hat{S}} Sim(s_i, s_j) \,]$$

The parameter λ helps control the redundancy (novelty) in the extracted summary. We use cosine similarity to calculate the similarity between two sentence embeddings. The query Q represents the case document by taking the average embeddings of all the sentences in the decision. Xiao and Carenini (2020) propose *MMR-Select* as an alternative approach to eliminate redundancy explicitly. It eliminates the greedy method, computing the query relevance to find suitable candidates with a neural model. *MMR* is more robust as it picks candidate sentences using the confidence scores, $P(y_i)$, produced by the neural model.

$$\text{MMR-Select} = arg \max_{s_i \epsilon D \setminus \hat{S}} [\, \lambda \, P(y_i) - (1-\lambda) \max_{s_j \epsilon \hat{S}} Sim(s_i, s_j) \,]$$

4.4 Redundancy Loss

The major limitation of explicit methods like *MMR* is the disconnect between the sentence scoring and sentence selection phases. Such techniques rely on the classifier to score the sentences in the document and check for redundancy later when selecting the final sentences for the summary. Thus, the classifier used to generate the confidence score does not implicitly learn how to handle redundancy. We can generate more informative summaries by teaching the neural model to avoid picking similar sentences. Xiao and Carenini (2020) propose adding a redundancy loss term L_{RD} to the cross-entropy loss function that penalizes the model for choosing two similar sentences with high confidence scores. The parameter β balances the importance we assign to the L_{CE} and L_{RD}. The neural models tend to classify more sentences as part of the summary for longer case documents. Therefore, we scale the redundancy loss L_{RD} defined by Xiao and Carenini (2020) to ensure that it does not explode as the length of the document increases, preventing it from overshadowing the cross-entropy loss.

$$L = \beta \, L_{CE} + (1-\beta) \, L_{RD}$$

$$L_{RD} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} P(y_i) \; P(y_j) \; Sim(s_i, s_j)$$

4.5 Extractive Summarization

4.5.1 Single-Task Models

We define extractive summarization as a binary classification problem where the proposed models decide whether a given sentence belongs to the fixed-length summary or not. We use the proposed models to generate the summary only for *Reasoning/Evidence* sentences as we can extract perfect matches for the other rhetorical role sentences (e.g., case issue and procedural background) by using regular expressions[7]. Our proposed models consist of two phases: Sentence Scoring and Sentence Selection, as shown in Figure 1. Initially, we use the approach explained in Section 4.1 to generate the embeddings for all the sentences in a given case document.

Sentence Scoring: Bidirectional Gated Recurrent Units (Bi-GRU) use two GRUs to simultaneously encode the sentence embeddings in both forward and backward directions. The concatenation of the forward and backward hidden states gives us the representation of each input sequence. A fully connected dense layer followed by the non-linear

[7] This is particular to the task of summarizing BVA decisions as introduced by (Zhong et al., 2019)

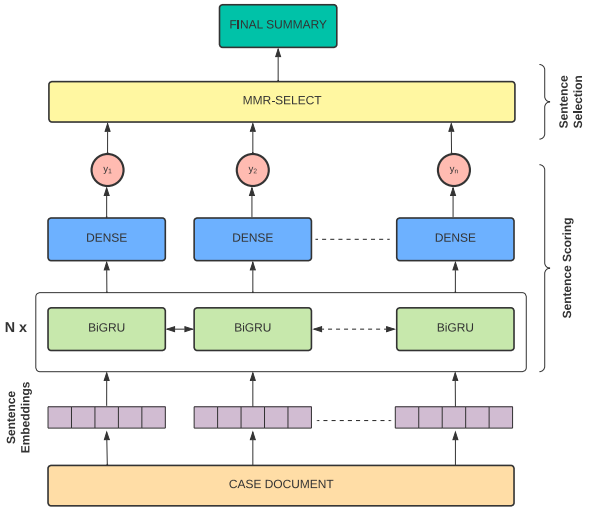
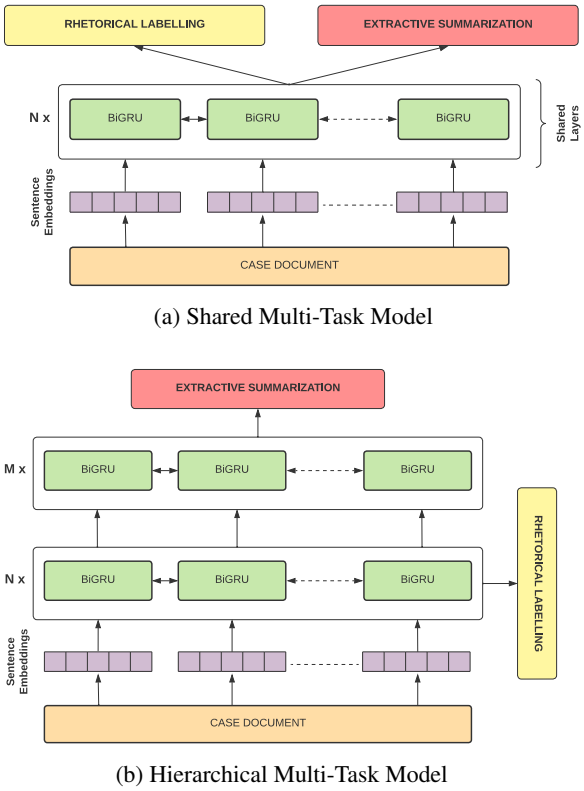


Figure 1: Architecture of the proposed Single-Task Extractive Summarization models: ST and ST+RdLoss



(a) Shared Multi-Task Model

(b) Hierarchical Multi-Task Model

Figure 2: Architecture of the proposed Multi-Task models: MT-Shared and MT-Hierarchical, respectively

softmax layer predicts the probability distribution over the two classes. We do not use transformer-based architectures for various reasons, including limited training data, input size limitations, and required computational resources.

Sentence Selection: *MMR*, discussed in Section 4.3, uses the sentence embeddings and confidence scores generated by the neural model to tackle redundancy explicitly and select the final sentences for the summary.

Accordingly, we refer to our single-task model described above as **ST**. We also propose another model, **ST+RdLoss**, to implicitly handle redundancy by using Redundancy Loss.

4.5.2 Multi-Task Models

Learning multiple tasks by jointly optimizing more than one criterion can help leverage the correlation between related tasks to improve performance (Liu et al., 2016; Ruder, 2017; Elnaggar et al., 2018). We propose using rhetorical role labeling as an auxiliary task to benefit the primary task of extractive summarization. We consider rhetorical role labeling a binary classification problem where the trained model learns to distinguish between *Reasoning/Evidence* and other rhetorical roles. This objective is similar to extractive summarization, where we only include sentences from *Reasoning/Evidence* in the summary.

Accordingly, we propose two multi-task models: **MT-Shared** and **MT-Hierarchical**, as shown in Figure 2. The first model shares the same Bi-GRU layers for both tasks. The second model follows a hierarchical order where rhetorical role labeling only uses lower-level Bi-GRU layers. The extractive summarization shares the lower-level Bi-GRU label with rhetorical labeling and has additional Bi-GRU layers on top to learn different features. Besides the shared Bi-GRU layers, each task has its task-specific layer comprising a fully connected dense layer followed by the non-linear softmax layer to predict the probability distribution over the two classes. Additionally, **MT-Shared+RdLoss** and **MT-Hierarchical+RdLoss** use Redundancy Loss discussed in Section 4.4.

5 Experiments

5.1 Baseline and Comparison

We evaluate[8] the performance of our proposed approaches with several baseline methods commonly used for extractive summarization. The two most common unsupervised methods are *MMR* and TextRank (Mihalcea and Tarau, 2004). TextRank uses a graph-based ranking system similar to PageRank to find relevant sentences. Since we are interested in keeping only Reasoning/Evidence sen-

[8]We do not compare the results with previous work by Zhong et al. (2019) due to the difference in the structure of the summaries generated and reproducibility issues.

tences for the summary, we also employ a binary classifier to filter out such sentences from the other rhetorical roles. We use CatBoost *(CB)* (Dorogush et al., 2018) and a GRU sequence labeler as the binary classifiers to identify the Reasoning/Evidence sentences [9]. The unsupervised methods then use the filtered-out sentences as the input to extract summaries. For *MMR*, we use the sentence embeddings discussed in Section 4.1 as the input, and cosine similarity measures the similarity between sentences. We generate the summaries for the *TextRank* approach using the Gensim[10] package that uses *BM25* scoring instead of TF-IDF or cosine similarity (Barrios et al., 2016). Thus we have four different baseline methods: **RL-CB+MMR** *(Cosine)*, **RL-GRU+MMR** *(Cosine)*, **RL-CB+TextRank** *(BM25)* and **RL-GRU+TextRank** *(BM25)*.

5.2 Implementation Details

We use five-fold cross-validation to find the best hyperparameters for our baseline and proposed models. The parameter λ for our *MMR*-based baseline models is determined using the training set described in Appendix D. We use hyperparameter tuning for *MMR* and Redundancy Loss to find the best values for λ and β, respectively. To train the multi-task learning models, we use the conventional cross-entropy loss function for the rhetorical role labeling task and alternate between the two datasets after every iteration. We randomly oversample the rhetorical role labeling task data to match the number of samples in the extractive summarization dataset. Therefore, in each mini-batch, we have training samples from either the extractive summarization or the rhetorical role labeling dataset and switch between them every other batch. We report the best set of hyperparameters corresponding to each model in Appendix E.

5.3 Evaluation Metrics

We use Recall to measure the performance for the binary classification problem. It measures how many sentences selected by the model are also part of the expert annotated summary. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) score counts the number of overlapping word sequences between candidate and reference summaries. ROUGE-1 and ROUGE-2 measure the unigram and bigram overlap, respectively. ROUGE-L finds the longest common subsequence matches to reflect sentence-level word ordering better.

[9] Our CatBoost model classifies one sentence at a time irrespective of the case document, whereas the GRU model takes all the sentences in the document as the input achieving an F1 score of **0.917** and **0.914**, respectively.

[10] https://radimrehurek.com/gensim/

6 Results and Analysis

6.1 Quantitative Evaluation

In Table 1, we report the number of sentences and tokens for the summaries in the test set generated by the different approaches. Our baseline and proposed models, which use MMR or MMR-Select, take the number of sentences as the input to generate the final summaries. Since the average number of sentences varies from 4.5 to 5.85 for the expert annotators, we choose n=5 as the suitable value for our models. For TextRank models, we require the number of tokens as the input, which we set to be 160 based on the statistics of the training set. Additionally, we observe that Annotator 2 tends to pick more sentences than the rest of the annotators resulting in the highest average token count of 171.3, while Annotator 1 and Annotator 3 pick fewer sentences on average. Our models' summaries have token counts similar to that of Annotator 4.

In terms of recall score for the binary classification problem, our proposed models demonstrate significant improvement compared to baseline methods, scoring more than twice the scores achieved by the baseline models (Appendix C). But, the recall metric score has limited use in evaluating the performance of the extractive summarization models as legal documents are often verbose, sometimes containing multiple sentences with similar meanings.

We compare the ROUGE scores for our annotators and models in Table 2. Annotator 3 has the highest ROUGE score among all the other annotators. In terms of ROUGE-1 and ROUGE-2, RL-CB+TextRank performs the best, while RL-CB+MMR scores the highest for ROUGE-L. Overall, CatBoost models perform better than GRU-based models. Also, we observe a sharp decline in ROUGE-2 scores compared to ROUGE-1 scores, indicating the baseline models' limited capability to pick the required sentences for the summary. We observe a general improvement in scores for the proposed approaches when adding the implicit redundancy measure, RdLoss, discussed in Section 4.4. Also, multi-task (MT) models tend to perform

	Sentences	Tokens	λ	β
Annotator 1	4.6 ± 1.46	130.55 ± 51.11	-	-
Annotator 2	5.85 ± 0.67	171.3 ± 44.43	-	-
Annotator 3	4.5 ± 0.889	137.5 ± 37.99	-	-
Annotator 4	5.55 ± 2.23	149.85 ± 79.44	-	-
RL-CB+MMR	5 ± 0	151.55 ± 33.76	0.9	-
RL-CB+TextRank	5.45 ± 1.31	153.2 ± 9.57	-	-
RL-GRU+MMR	5 ± 0	156.6 ± 38.81	0.9	-
RL-GRU+TextRank	5.05 ± 1.23	158.7 ± 13.24	-	-
ST	5 ± 0	156.2 ± 35.48	0.8	-
ST+RdLoss	5 ± 0	149.35 ± 33.92	0.9	0.85
MT-Shared	5 ± 0	150.8 ± 27.2	0.6	-
MT-Shared+RdLoss	5 ± 0	157 ± 34.1	0.9	0.775
MT-Hierarchical	5 ± 0	151.75 ± 28.04	0.9	-
MT-Hierarchical+RdLoss	5 ± 0	154.75 ± 36.42	1	0.9

Table 1: Overview of the **number of sentences** and **tokens** in the summaries generated by expert annotators and models. Values of parameters, λ and β, for *MMR*, *MMR* and RdLoss-based approaches.

	Annotator 1			Annotator 2			Annotator 3			Annotator 4		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Annotator 1	100	100	100	73	60.7	61.4	64.8	52.8	54.4	62.6	52	54.9
Annotator 2	59.9	48.9	49.8	100	100	100	63.1	54.3	54.9	55.3	43.5	45.4
Annotator 3	65.9	54	55.3	**79.3**	**69.1**	**69.7**	100	100	100	68.4	56.7	59.7
Annotator 4	65	54	56.9	70.5	56.3	58.8	70.6	58.2	61.6	100	100	100
RL-CB+MMR	52.2	32	36.8	54.8	37.8	40.9	56.8	36.8	39.6	53	32.8	36.6
RL-CB+TextRank	53	33.7	34	56.7	40.4	34.9	57.3	36.8	36.7	55.4	37.4	34.3
RL-GRU+MMR	51.9	32.1	36.5	52.9	34.6	38.5	53.2	31.4	36.4	51.3	30.2	36.1
RL-GRU+TextRank	51.5	31.6	33	56.7	40	33.4	54.8	32.8	33.9	49.6	27.3	28.6
ST	67.3	55.2	57.8	64.6	51.4	52.1	73.6	60.7	62.5	65.2	51.4	55.1
ST+RdLoss	68	57.6	59.7	63.2	51	51.4	75.3	64.8	65.7	66.8	54.8	57.8
MT-Shared	70.6	60.1	60.5	64.3	52.2	51.9	74.1	61.9	62.8	66.9	54.6	56.8
MT-Shared+RdLoss	70	59.2	59.7	65	52.3	52.3	**77.3**	**67.4**	**68.3**	**68.8**	58.1	60
MT-Hierarchical	70.6	59.9	60.9	63.3	50.6	51.6	77.1	66.5	66.5	68.1	57.1	58.2
MT-Hierarchical+RdLoss	**71**	**60.5**	**63.1**	64.4	52.1	53.5	75.4	64.3	65.6	68.6	**58.3**	**62**

Table 2: **ROUGE** scores averaged for decisions in test set and compared to the four expert annotators. Best score for annotators, baseline approaches, and proposed models are highlighted in orange, yellow, and green, respectively.

better than single-task (ST) models. Overall, MT-Shared+RdLoss and MT-Hierarchical+RdLoss perform the best, scoring higher than three expert annotators. Our models fall short for Annotator 2 as they have annotated more sentences and thus have more extended summaries than ones generated by our trained models.

We also present the values of the parameter λ used in *MMR* and *MMR* in Table 1. The models that employ the implicit redundancy check (RdLoss) tend to have higher values for λ than those with just an explicit redundancy check, indicating that the additional term in the loss function helps the model tackle redundancy better.

We also compare the performance in terms of the outcome of the legal decisions in Figure 3. The findings with the remanded outcome have higher scores, constituting approximately 50% of the train and test data. MT-Shared+RdLoss per-

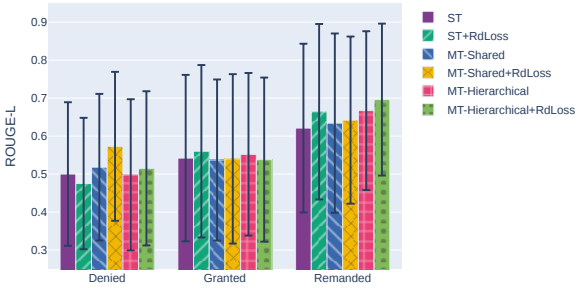


Figure 3: Comparison of the performance for different decision outcomes in terms of ROUGE-L.

	Denied		Granted		Remanded		Total	
	Rank	Ad.	Rank	Ad.	Rank	Ad.	Rank	Ad.
Annotator 3	2.83 ± 1.17	0.67	3.75 ± 1.89	0.50	2.9 ± 1.37	**1.00**	3.05 ± 1.39	**0.8**
Annotator 4	4.17 ± 0.75	0.50	2.5 ± 1.0	**0.75**	2.9 ± 1.73	0.90	3.2 ± 1.47	0.75
MT-Hierarchical+RdLoss	**1.83 ± 1.6**	**0.83**	2.5 ± 1.91	**0.75**	3.3 ± 1.57	0.60	2.7 ± 1.69	0.7
MT-Shared+RdLoss	2.0 ± 1.26	**0.83**	2.25 ± 1.89	0.50	2.3 ± 1.42	0.80	2.2 ± 1.4	0.75
ST+RdLoss	2.5 ± 1.64	0.67	**2.0 ± 1.41**	**0.75**	**1.8 ± 1.03**	0.80	**2.05 ± 1.28**	0.75

Table 3: Qualitative analysis of the summaries in terms of **Rank** (lower is better) and **Adequacy** (higher is better). Best score for annotator and proposed models are highlighted in orange and green, respectively.

forms notably better for denied outcomes, while MT-Hierarchical+RdLoss scores the highest for remanded decisions. The shared layers between the primary and auxiliary tasks combined with a supplementary dataset consisting of only granted and denied findings make it difficult for the MT-Shared and MT-Shared+RdLoss to generate better summaries for remand cases. Also, remanded cases can differ in terms of rhetorical role distribution from granted and denied decisions, as they end in the BVA sending the case back to the regional VA office, often instructing it extensively on what to do next. In contrast, the hierarchical multi-task (MT) models leverage the additional GRU layer to perform better for such outcomes. As shown in Figure 4, we note a significant decrease in performance for lengthy decisions for all the methods that could be attributed both to the limitations of the GRU models and insufficient training data for such cases.

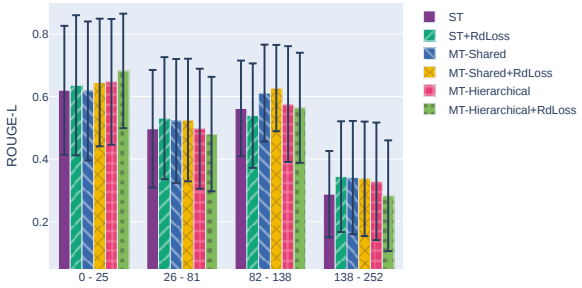


Figure 4: Performance for different methods based on the number of sentences in the decisions in terms of ROUGE-L.

6.2 Qualitative Evaluation

We further perform manual qualitative analysis to better understand the proposed approaches' performance compared to the expert annotators. We generated the summaries for each decision in the test set using three of the proposed models and compared them with the ones from two expert annotators. The outputs were randomized, and a fifth annotator with expertise in the legal domain (the third author) ranked each summary and checked if it was adequate against the two human-drafted reference summaries. Following (Zhong et al., 2019), we consider a summary adequate if it identifies all major legal issues and resolutions in the case. We then rank summaries based on the additional information they contain about the case narrative and their coherence. We assign the same rank to two or more summaries if they are duplicate, near-duplicate, or semantically equivalent.

We report the results for qualitative analysis in Table 3. Overall our proposed models almost always ranked better than the two annotators. ST+RdLoss ranked best on average and for decisions with the outcome as remanded or granted. MT-Hierarchical+RdLoss ranked better for the denied decisions. In terms of adequacy, both Annotator 3 and MT-Hierarchical+RdLoss achieved reasonable accuracy. The multi-task (MT) models achieve higher ROUGE scores but fail to produce summaries of good qualities for remand cases. A possible explanation is that the design of the annotation type system is more suitable for evidence-based findings (e.g., an in-service stressor has caused a particular disability of the veteran), which are most clearly present in denied and granted cases. Remand cases challenge this annotation type system because the BVA determines that we cannot make a finding based on the current evidence. Summary annotators must cope with this slight semantic mismatch and may develop an individual lexical bias in assigning sentence types. The MT-Hierarchical+RdLoss model can potentially overfit that bias because it contains an additional GRU layer unaffected by the rhetorical role supervision signal.

7 Discussion

Our proposed approaches generate higher-scoring summaries than baseline methods and expert annotators in terms of ROUGE scores and appear to be competitive in a qualitative expert ranking

evaluation. The domain-specific pre-trained transformers help produce sentence embeddings capable of capturing the semantics of legal decisions in a way conducive to being used as a component in our proposed setup for summary generation. We can use these embeddings to train simple models like GRU effectively. Adding explicit methods like *MMR* helps tackle redundancy to generate more informative summaries even for verbose legal decisions. We further improve the performance by using the weighted cross-entropy loss function combined with redundancy loss (*RdLoss*). The additional loss term helps train models that can handle redundancy implicit.

The quantitative measures show the effectiveness of both the single-task and multi-task models, even with a limited dataset containing just 120 legal decisions. The supplementary dataset used to train the same model for the rhetorical role labeling task helps learn better representations and achieve better summary ROUGE scores. Such correlated tasks and datasets prove helpful in low-resource settings such as ours to improve the performance further at no additional annotation costs, assuming they stem from the same legal domain. Further qualitative analysis with the help of an expert annotator shows our proposed approaches rank at least as good as human annotators. It also indicates the need for better quantitative metrics to evaluate the quality of the summaries. The specialized BVA domain we experiment in is relatively narrow in scope and highly regular in its document structure. Our proposed methods seem promising as they work well with limited annotated datasets and computational resources but warrant further investigation and validation on larger, more diverse datasets.

8 Conclusion

Training models that can generate extractive summaries of legal opinions comparable to expert annotators in a low-resource setting can be challenging. We demonstrate that domain-specific pre-trained transformers and multi-task training with rhetorical role labeling can effectively train sequential extractive summarizaton models (in our case, GRUs) on a relatively constrained domain of cases. The proposed methods implement implicit and explicit redundancy checks to maximize the information and minimize the redundancy in summaries. In our experiments, we systematically analyze the performance of different techniques, both quantitatively and qualitatively. The results verify the efficacy of our model design. In the future, we plan to extend our work to other decision types and jurisdictions in the legal domain. We further plan to explore the discrepancy between our quantitative results favoring multi-task models and qualitative evaluation preferring summaries by single-task architectures.

Limitations

We only consider two different rhetorical roles for summarization. Decisions from other subdomains and jurisdictions might require the inclusion of more rhetorical functions in the final summary. A suitable domain-specific pre-trained transformer might not be available to produce the necessary sentence embeddings. In such cases, we would have to rely on conventional approaches like Universal Sentence Encoder (USE) (Cer et al., 2018), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov et al., 2013). Our methods do not automatically scale well for very long decisions, so we must ensure ample availability of such decisions in the training set.

The BVA decisions we use have a relatively regular structure and are constrained to cases deciding issues of compensation for service-connected PTSD disabilities of veterans, which is only a subset of the issues adjudicated by the BVA. The decisions discuss similar aspects of medical diagnoses, stressors experienced in service, and causation. We still need to validate if we could extend our proposed approach to collections of cases that include more diverse legal issues and fact patterns. Legal texts often contain language that looks similar on the surface but is different in its semantics and vice versa. More complex textual phenomena may challenge the redundancy-focused components of the system.

Ethical Concerns

The Board of Veterans' Appeals publicly releases its decisions (including the ones in our datasets) on its publicly available website. Generally, in BVA decisions, the veteran is not named explicitly. While the nature of disability compensation claims (including those relating to PTSD) is sensitive, we chose these particular datasets because of several factors: (a) the scarcity of expert-annotated legal decisions in the public domain was suitable for summarization research when conducting the experiments; (b) the availability of PTSD-related annotated decisions from the Research Laboratory

for Law, Logic, and Technology (LLT Lab) at Hofstra University with a matching set of summaries and (c) our prior experience where we worked with BVA decisions and U.S. Veterans Law. Additionally, proving the requirements of a service-connected PTSD disability using relevant evidence is legally sufficiently complex to provide a suitable testbed to evaluate the proposed summarization techniques. At the same time, it is a reasonably closed domain such that the comparative ranking of candidate summaries is more straightforward and coherent.

The biases, inequalities, and under-representations encoded in the pre-trained transformer models might get inherited by our GRU models and propagated to the generated summaries (Bommasani et al., 2020). To deploy these models in a production system, one must thoroughly check for such biases by comprehensively evaluating summarization performance across relevant groups (e.g., gender and race) using tests such as, for example, the recently proposed Pronoun-Ranking Test (Silva et al., 2021).

An automatic summarization model for legal decisions may perform worse for some partitions of its domain than others. For example, in the BVA context, cases about rarely occurring disabilities or special legal and military situations may lead to worse summaries because of sparseness in the training data. It might disparately affect groups that are supposed to be treated equally if group membership tends to coincide with such configurations. If lawyers use (and potentially depend on) automatic summarization tools to assist clients, screening such systems may become necessary. For example, one can engage domain experts to curate datasets with better representation across different types of injuries and legal phenomena that might be uncommon or related to particular groups. Still, the quantitative improvement of additional development data obtained toward more consistent summary quality may be uncertain and model-dependent. At the very least, it will help reveal performance disparities and increase expert user awareness around the limitations of automatic summarization technology in the legal domain.

9 Acknowledgements

We want to thank Rashid Haddad and Santosh T.Y.S.S. from the Technical University of Munich for their assistance in analyzing the results and feedback about the manuscript.

References

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems*, pages 3–12. IOS Press.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2021a. Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, pages 1–38.

Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021b. Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Ahmed Elnaggar, Christoph Gebendorfer, Ingo Glaser, and Florian Matthes. 2018. Multi-task deep learning for legal document translation, summarization and multi-label classification. In *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*, pages 9–15.

Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining different summarization techniques for legal text. In *Proceedings of the workshop on innovative hybrid approaches to the processing of textual data*, pages 115–123.

Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E Ho, Mark S Krass, and Matthias Grabmair. 2021. Context-aware legal citation recommendation using deep learning. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 79–88.

Mi-Young Kim, Ying Xu, and Randy Goebel. 2012. Summarization of legal texts with high cohesion and automatic compression rate. In *JSAI International Symposium on Artificial Intelligence*, pages 190–204. Springer.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chao-Lin Liu and Kuan-Chun Chen. 2019. Extracting the gist of chinese judgments of the supreme court. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 73–82.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. IJCAI'16, page 2873–2879. AAAI Press.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. 2017. A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In *JURIX*, pages 125–134.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

M Saravanan and Balaraman Ravindran. 2010. Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artificial Intelligence and Law*, 18(1):45–76.

Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues*, 58:21.

Jaromir Savelka, Hannes Westermann, Karim Benyekhlef, Charlotte S Alexander, Jayla C Grant, David Restrepo Amariles, Rajaa El Hamdani, Sébastien Meeùs, Aurore Troussel, Michał Araszkiewicz, et al. 2021. Lex rosetta: transfer of predictive models across languages, jurisdictions, and legal domains. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 129–138.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389.

Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. In *ASAIL@ ICAIL*.

Wen Xiao and Giuseppe Carenini. 2020. Systematically exploring redundancy reduction in summarizing long documents. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 516–528, Suzhou, China. Association for Computational Linguistics.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In *Legal Knowledge and Information Systems*, pages 33–42. IOS Press.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168.

Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D Ashley, and Matthias Grabmair. 2019. Automatic summarization of legal decisions using iterative masking of predictive sentences. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 163–172.

A Inter-Annotator Agreement

We present the pairwise inter-annotator agreement score using the Cohen-Kappa coefficient in Table 4. The scores vary from 0.46 to 0.55, indicating low agreement among annotators. However, this metric is not ideal, as annotators can mark up different sentences that still address similar aspects of the case. We also calculate the pairwise metric for our proposed methods in Table 5 to measure how similar the outputs are. The highest agreement is between MT-Shared+RdLoss and MT-Hierarhical+RdLoss. The summaries generated by these two approaches mostly differ by 1 or 2 sentences.

B Sentence Embeddings Visualization

We illustrate the T-SNE projection of the sentence embeddings for the 50 decisions in the rhetorical labeling dataset in Figure 5. The sentence embeddings can easily separate annotation types like Citation, LegalRule, and Finding sentences, as these classes have a somewhat unique vocabulary. However, the embeddings for other annotation types do not have such a clear distinction and overlap.

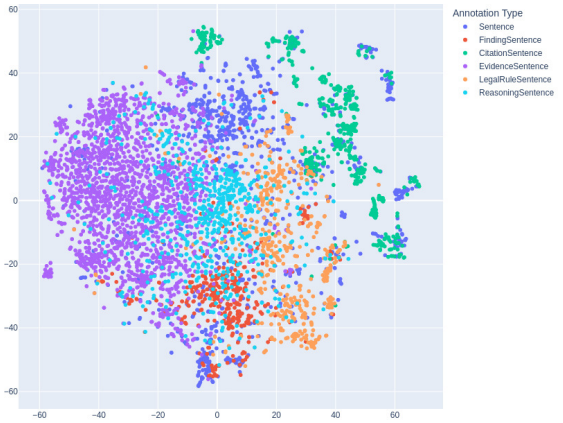


Figure 5: T-SNE projection of the 768-dimensional sentence embeddings to two-dimensional space for the 6984 sentences and corresponding rhetorical roles.

C Additional Results

For the Recall score presented in Table 6, the baseline model, RL-CB+TextRank, performs the best for three of the four annotators, while RL-CB+MMR scores the highest for the remaining annotator. Our proposed models perform notably better than the baseline models. Overall, the multi-task (MT) based models outperform the single-task (ST) models. MT-Hierarchical+RdLoss performs the best for Annotator 1 and Annotator 4.

	Annotator 1	Annotator 2	Annotator 3	Annotator 4
Annotator 1	100	**49.4**	46.1	46.7
Annotator 2	49.4	100	**54.4**	46.7
Annotator 3	46.1	**54.4**	100	49.4
Annotator 4	46.7	46.7	**49.4**	100

Table 4: Cohen Kappa Score for the four annotators

	ST	ST+RdLoss	MT-Shared	MT-Shared+RdLoss	MT-Hier	MT-Hier+RdLoss
ST	100	73.5	**74.4**	70.2	65.7	63.8
ST+RdLoss	73.5	100	**75.5**	73.3	71.2	72.8
MT-Shared	74.4	75.5	100	**79.7**	73.3	72.3
MT-Shared+RdLoss	70.2	73.3	79.7	100	77.9	**81.5**
MT-Hier	65.7	71.2	73.3	77.9	100	**81.2**
MT-Hier+RdLoss	63.8	72.8	72.3	**81.5**	81.2	100

Table 5: Cohen Kappa Score for the six proposed methods

	Annotator 1	Annotator 2	Annotator 3	Annotator 4
RL-CB+MMR	0.255 ± 0.265	0.308 ± 0.272	0.283 ± 0.261	0.231 ± 0.263
RL-CB+TextRank	0.285 ± 0.29	0.35 ± 0.247	0.276 ± 0.25	0.294 ± 0.24
RL-GRU+MMR	0.227 ± 0.218	0.25 ± 0.22	0.215 ± 0.236	0.191 ± 0.273
RL-GRU+TextRank	0.243 ± 0.282	0.333 ± 0.286	0.247 ± 0.279	0.18 ± 0.248
ST	0.476 ± 0.257	0.45 ± 0.254	0.52 ± 0.25	0.422 ± 0.263
ST+RdLoss	0.499 ± 0.267	0.467 ± 0.251	0.578 ± 0.233	0.457 ± 0.273
MT-Shared	0.498 ± 0.242	**0.475 ± 0.225**	0.528 ± 0.238	0.441 ± 0.237
MT-Shared+RdLoss	0.486 ± 0.238	0.45 ± 0.23	0.58 ± 0.236	0.466 ± 0.237
MT-Hierarchical	0.508 ± 0.272	0.458 ± 0.259	**0.584 ± 0.255**	0.467 ± 0.254
MT-Hierarchical+RdLoss	**0.519 ± 0.274**	0.467 ± 0.251	0.544 ± 0.239	**0.495 ± 0.265**

Table 6: **Recall** scores for the extractive summarization task averaged for all the decisions in the test set. Best score for baseline and proposed models are highlighted in yellow and green, respectively.

D Hyperparameter Tuning for Baseline Methods

This section briefly discusses how we determine the parameter λ for *MMR* in the baseline methods: RL-CB+MMR and RL-GRU+MMR. Once we have filtered out the Evidence/Reasoning sentences using the CatBoost or GRU classifier, we generate the summaries using different values of λ for all the decisions in the training set. We also vary the number of sentences to determine the optimal length for the summary. We then measure the impact of λ and summary length on the total number of words in the summary, recall score, and ROUGE-L, as shown in Figure 6. We choose values that result in the highest recall and ROUGE-L but consist of the number of tokens comparable to experts and proposed methods. The same values are used to produce the summaries for decisions in the test set.

E Hyperparameters tuning for Proposed Methods

We use a combination of random and bayesian searches to find the best set of hyperparameters for our models. We use the random search to find an approximate search space suitable for our model, followed by a more targeted search using bayesian optimization. All our models achieve the best scores using just one GRU layer. Also, since we use sentence embeddings derived from pre-trained transformers, our models converge quickly with a minimal number of epochs. We report the best set of hyperparameters for all our models in Table 7

In Figure 7, we demonstrate the effect of λ on our proposed methods for the training set. The number of tokens and recall score increase linearly with the value of λ up to 0.5 and then changes very slowly. Assigning less weight to λ forces the model to pick very dissimilar sentences and results in poor performance. Therefore, we set the minimum value of λ as 0.5 for hyperparameter tuning to find the right balance between similarity and redundancy. Additionally, the λ has a minimal impact on baseline methods.

F Examples

Based on the assessment of our qualitative analysis, we demonstrate the summaries generated by annotators and models for two decisions, one each

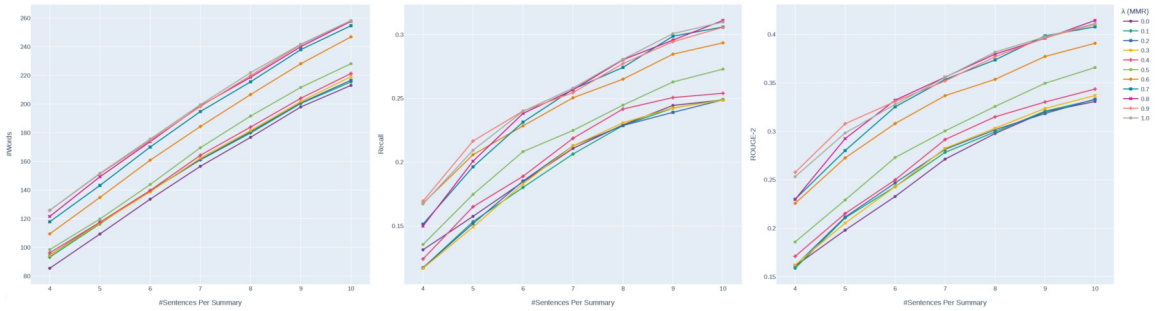


Figure 6: We estimate the best value for the parameter λ required for *MMR* by varying it from 0 to 1 at intervals of 0.1 for the training set.

	num_layers	hidden_size	dropout (RL)	dropout (ES)	batch_size	epochs	learning_rate
ST	1	128	-	0.5	8	5	0.00261
ST+RdLoss	1	64	-	0.6	8	9	0.00441
MT-Shared	1	512	0.4	0.5	4	6	0.00019
MT-Shared	1	512	0.6	0.4	4	11	0.00018
MT-Hier	1 + 1	128 + 512	0.5	0.6	8	5	0.00143
MT-Hier+RdLoss	1 + 1	128 + 256	0.6	0.4	4	8	0.00053

Table 7: Final hyperparameters required to train the proposed extractive summarization models.

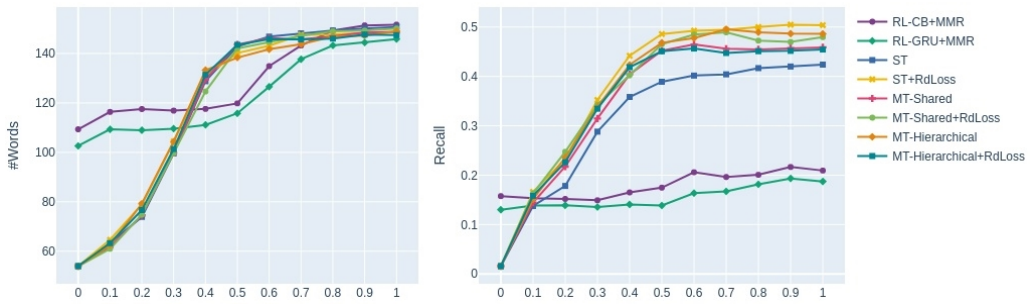


Figure 7: Effect of varying λ for different models on training set in terms of number of words in summary, extractive Recall and ROUGE-L scores

from denied and remanded outcomes (ordered by their rank). The multi-task models generate better summaries than the two annotators in the example shown in F.1. For the second example in F.2, the annotators are ranked better than all the proposed methods, which fail to produce adequate summaries.

F.1 Denied Example

MT-Hierarchical+RdLoss:

"Furthermore, Vietnam casualty logs do not confirm the supposed death of PL nor do the deck logs record PLs unavailability for duty due to injury or death. Overall, the veterans report of stressors while serving on the USS CORAL SEA lack any indicia of reliability or credibility. Certainly, there is no credible supporting evidence of these alleged stressors. Based upon the above, the Board finds that the preponderance of the evidence establishes that the veteran was not engaged in combat, and that his allegations of exposure to combat and non-combat stressors are not credible. The Board, therefore, finds that the veterans PTSD is not shown to have its origins in his military service."

MT-Shared+RdLoss:

"The Board also notes that the veteran appears to have recanted his original assertion, contained in a March 1995 VA mental health clinic visit, that he "frequently ran into combat" on mainland visits. Overall, the veterans report of stressors while serving on the USS CORAL SEA lack any indicia of reliability or credibility. Certainly, there is no credible supporting evidence of these alleged stressors. Based upon the above, the Board finds that the preponderance of the evidence establishes that the veteran was not engaged in combat, and that his allegations of exposure to combat and non-combat stressors are not credible. The Board, therefore, finds that the veterans PTSD is not shown to have its origins in his military service."

Annotator 4:

"The veteran's service medical records do not reflect his report of exposure to combat or non-combat stressors. Overall, the veteran's report of stressors while serving on the USS CORAL SEA lack any indicia of reliability or credibility. Based upon the above, the Board finds that the preponderance of the evidence establishes that the veteran was not engaged in combat, and that his allegations of exposure to combat and non-combat stressors are not credible. The Board, therefore, finds that the veteran's PTSD is not shown to have its origins in his military service."

Annotator 3:

"The veteran's service medical records do not reflect his report of exposure to combat or non-combat stressors. There is no report that an individual named PL was unavailable for duty due to injury or death. The Board, therefore, finds that the veteran's PTSD is not shown to have its origins in his military service. The claim for service connection for PTSD, therefore, must be denied."

ST+RdLoss:

"Overall, the veterans report of stressors while serving on the USS CORAL SEA lack any indicia of reliability or credibility. Certainly, there is no credible supporting evidence of these alleged stressors. Based upon the above, the Board finds that the preponderance of the evidence establishes that the veteran was not engaged in combat, and that his allegations of exposure to combat and non-combat stressors are not credible. The Board, therefore, finds that the veterans PTSD is not shown to have its origins in his military service. As the veteran was not engaged in combat, the evidentiary presumptions contained in 38 U.S.C.A. § 1154(b) and 38 C.F.R. § 3.304(f) are not applicable in this case."

F.2 Remand Example

Annotator 4:

"The remand instructions stated that the examiner was to render an opinion as to whether the documented in-service headaches and fainting spells were manifestations of her diagnosed acquired psychiatric disorder (to include PTSD, depression, and anxiety). Regarding the Veterans complaints of headaches and fainting spells in service, the examiner stated, "The Veteran reports multiple pain issues including headaches and back pain. This, however, does not respond to the question as to whether these headaches and fainting spells were manifestations of a current psychiatric disorder. In addition, the Veteran, through her representative, has asserted that her PTSD and depression are caused or aggravated by her service connected migraines. As the December 2016 opinion does not clearly address this issue, an additional medical opinion is warranted to determine whether the Veterans PTSD is proximately due to or aggravated by her service-connected migraines."

Annotator 3:

"The examiner noted mental health diagnoses of

chronic PTSD with secondary generalized anxiety disorder and major depressive disorder. Regarding the Veterans complaints of headaches and fainting spells in service, the examiner stated, "The Veteran reports multiple pain issues including headaches and back pain. This, however, does not respond to the question as to whether these headaches and fainting spells were manifestations of a current psychiatric disorder. In addition, the Veteran, through her representative, has asserted that her PTSD and depression are caused or aggravated by her service connected migraines. The examiner must also opine as to whether it is at least as likely as not that the in-service episodes of fainting and headaches were a manifestation of a currently diagnosed acquired psychiatric disorder."

MT-Shared+RdLoss:

"The remand instructions stated that the examiner was to render an opinion as to whether the documented in-service headaches and fainting spells were manifestations of her diagnosed acquired psychiatric disorder (to include PTSD, depression, and anxiety). The examiner was also advised to address the medical literature in the October 2016 Appellate Brief suggesting that fainting and headaches can be physical symptoms of PTSD and can occur as a result of exposure to trauma and the Veterans September 1980 in-service reports of headaches and fainting spells. Regarding the Veterans complaints of headaches and fainting spells in service, the examiner stated, "The Veteran reports multiple pain issues including headaches and back pain. As the December 2016 opinion does not clearly address this issue, an additional medical opinion is warranted to determine whether the Veterans PTSD is proximately due to or aggravated by her service-connected migraines. The examiner must also opine as to whether it is at least as likely as not that the in-service episodes of fainting and headaches were a manifestation of a currently diagnosed acquired psychiatric disorder."

ST+RdLoss:

"The Veteran was afforded an additional VA psychiatric evaluation in December 2016. The examiner noted mental health diagnoses of chronic PTSD with secondary generalized anxiety disorder and major depressive disorder. Regarding the Veteran's complaints of headaches and fainting spells in service, the examiner stated, "The Veteran reports multiple pain issues including headaches and back pain. These issues could be exacerbated by emotional distress but are not directly related." This, however, does not respond to the question as to whether these headaches and fainting spells were manifestations of a current psychiatric disorder. This statement requires clarification. In addition, the Veteran, through her representative, has asserted that her PTSD and depression are caused or aggravated by her service connected migraines. See June 2017 Appellate Brief. As the December 2016 opinion does not clearly address this issue, an additional medical opinion is warranted to determine whether the Veteran's PTSD is proximately due to or aggravated by her service-connected migraines."

MT-Hierarchical+RdLoss:

" The Veteran was afforded an additional VA psychiatric evaluation in December 2016. This, however, does not respond to the question as to whether these headaches and fainting spells were manifestations of a current psychiatric disorder. In addition, the Veteran, through her representative, has asserted that her PTSD and depression are caused or aggravated by her service connected migraines. See June 2017 Appellate Brief. As the December 2016 opinion does not clearly address this issue, an additional medical opinion is warranted to determine whether the Veteran's PTSD is proximately due to or aggravated by her service-connected migraines. Obtain a VA addendum opinion by the same December 2016 examiner, (or another appropriate examiner if unavailable), to provide opinions as to whether it is at least as likely as not (50 percent or better probability) that the Veteran's PTSD was caused by OR aggravated (i.e., permanently worsened beyond the natural progress of the disorder) by her service-connected migraine headaches."